FEB 2024

Data Mining Final

Adela Cho, Mike Meissner, Joey Gaule, Joseph Strickland

Setting the Stage

01

Analysis Plan

01 Goals of Analysis

02 Methodological

03 Sampling Methods used for Data Collection

04 Data Description

05 Expected Results

Stakeholder:

Chief Marketing Officer

Our team hopes to provide a nuanced understanding of your customer base to provide specific product recommendations and comprehensive marketing plans that the CMO could execute

Sampling Method

- Collected from grocery firm’s database by Prof. Omar Rome Hernandez with the Haas School of Business at UC Berkeley

BerkeleyHaas
Haas School of Business
University of California Berkeley

Data Description

- Each observation is an aggregate of individual customer’s purchases

People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company’s website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company’s website in the last month

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Goal

Where We Are

- Strong presence in market

Where We Want to Be & Plan

- Lack data-driven customer segmentation, hindering targeted marketing strategies & optimal resource allocation
- Address gap by employing k-means clustering to identify distinct customer segments with unique characteristics
- Utilize linear regression to develop marketing plan tailored to each segment, maximizing campaign effectiveness & ensuring alignment with specific customer needs

Expected Results

- Obtain customer segments backed by data
- Create comprehensive marketing plans for each customer segment

EXECUTIVE SUMMARY

In the examination of data collected from a grocery firm’s database, our team identified 4 separate clusters in the consumer base utilizing K-Means clustering. The 4 clusters we identified were “The Affluent Connoisseurs”, “Budget-Focused Digital-Savvy Young Parents”, “Upper Mid-Level Affluents”, and “The Economical Engagers” and provided tailored marketing strategies such as creating a curated wine selection, value packs, exclusive online wine tastings, or bulk and economy buys.

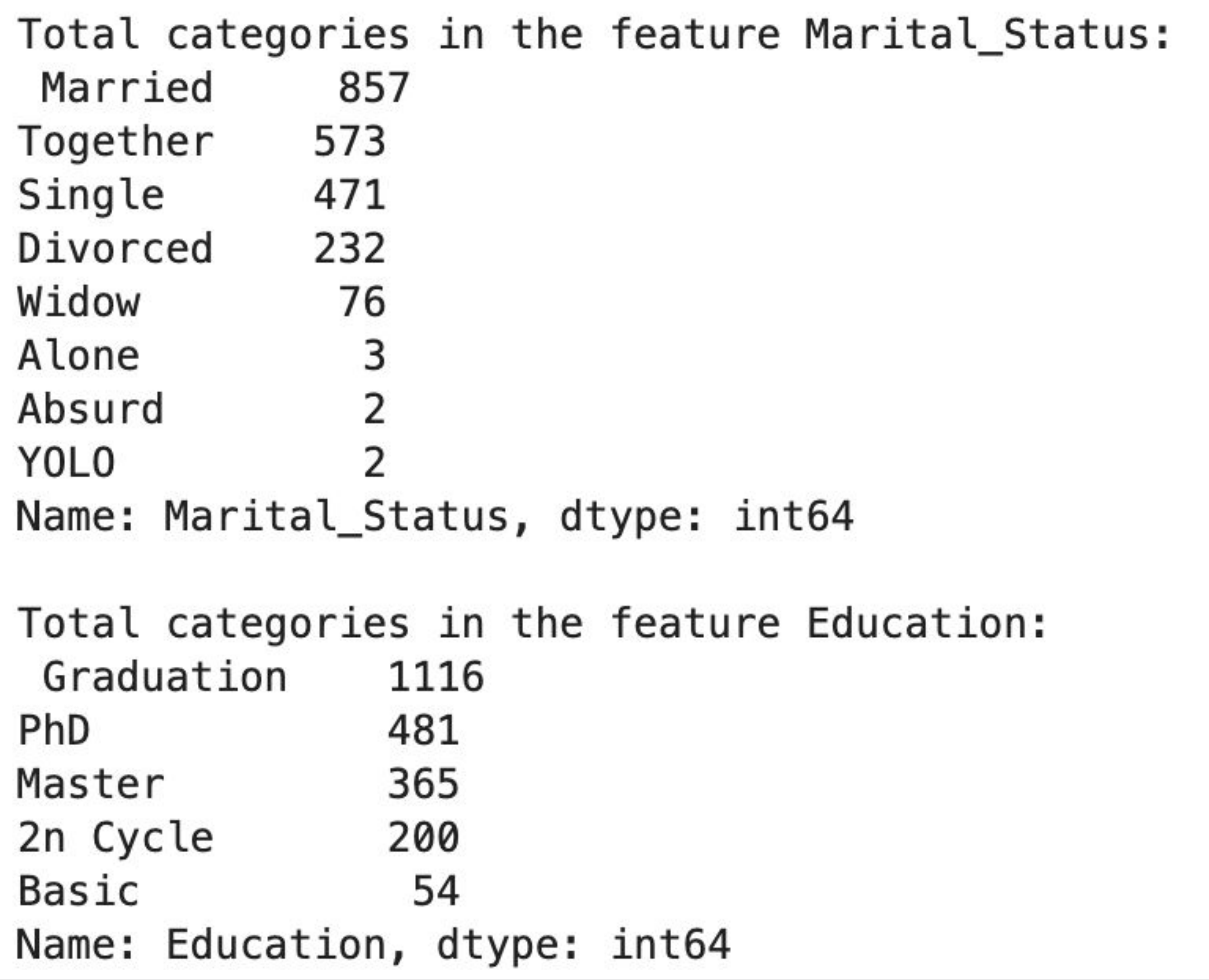
Analysis & Methodology

02

Model Choice

- ***Unsupervised: K-means clustering***
  - Assists in grouping dataset into distinct, non-overlapping clusters
  - Discovers natural groupings in data and help identify anomalies in the dataset
- ***Supervised: Multiple linear regression***
  - Allows for multiple predictors and analysis of strength and type of relationship between variables

EDA



```
Total categories in the feature Marital_Status:
 Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
Absurd        2
YOLO          2
Name: Marital_Status, dtype: int64

Total categories in the feature Education:
 Graduation    1116
PhD            481
Master         365
2n Cycle       200
Basic           54
Name: Education, dtype: int64
```

Figure 1) Analysis of Categorical Features. Total Levels in "Marital Status" and "Education"

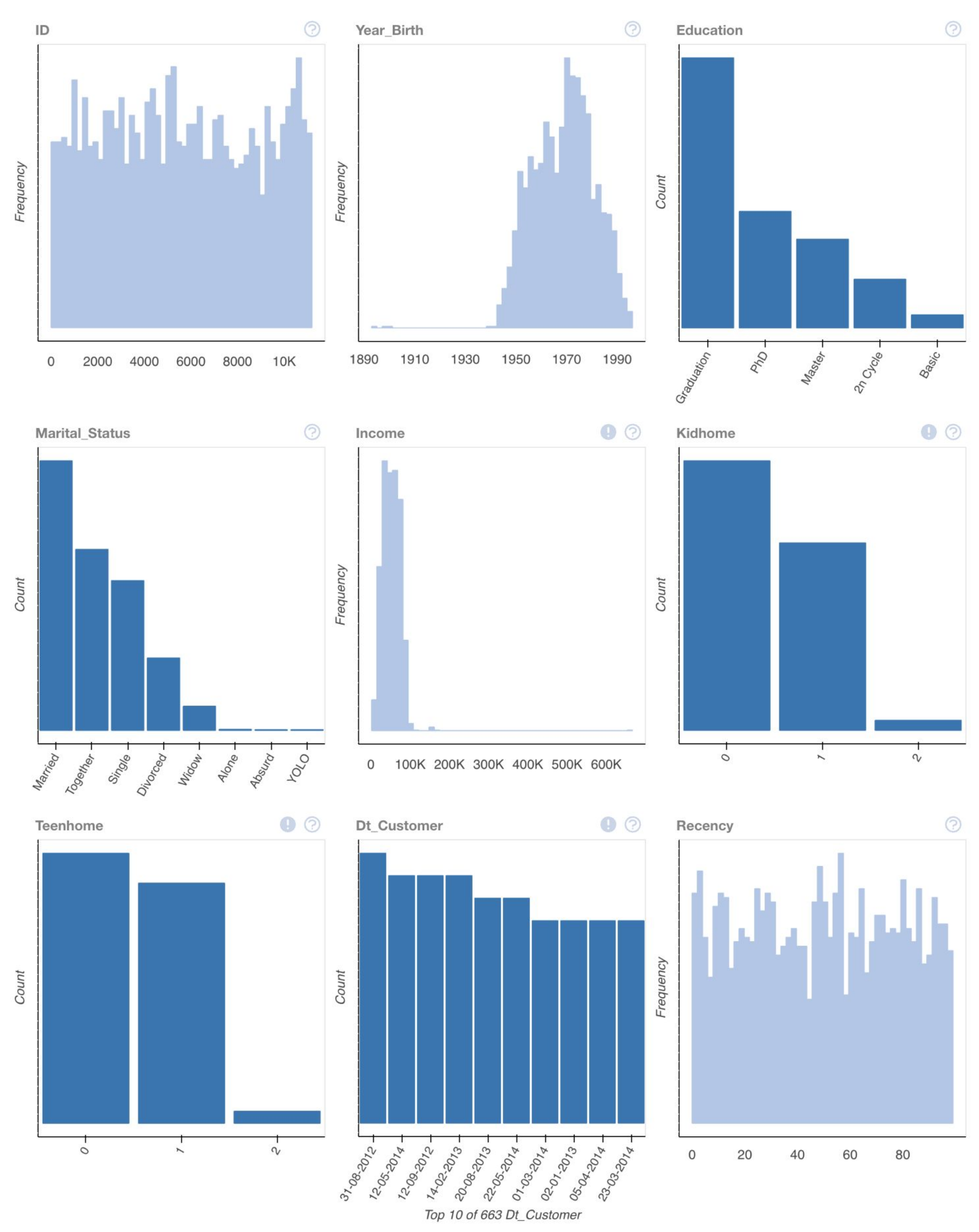
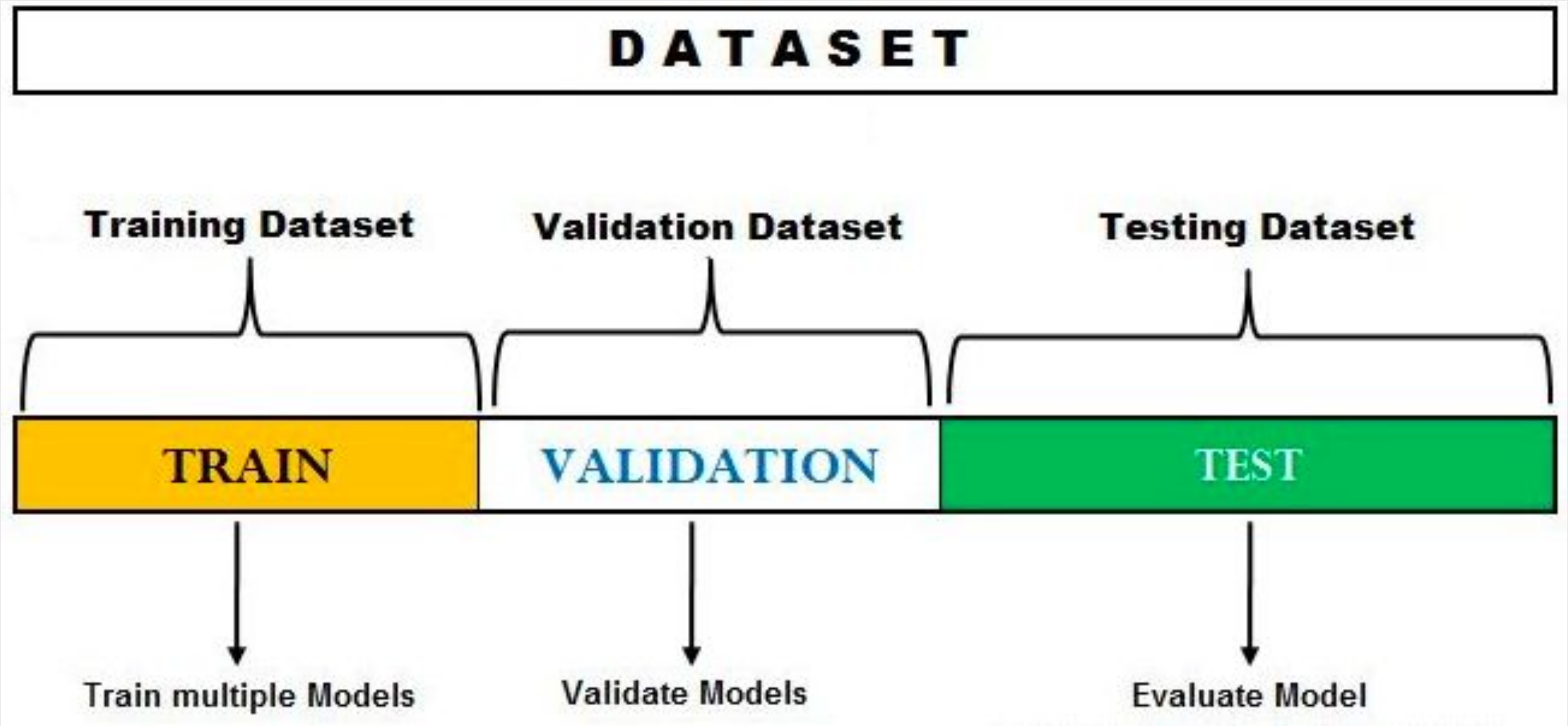


Figure 2) Count or Frequency of Variables

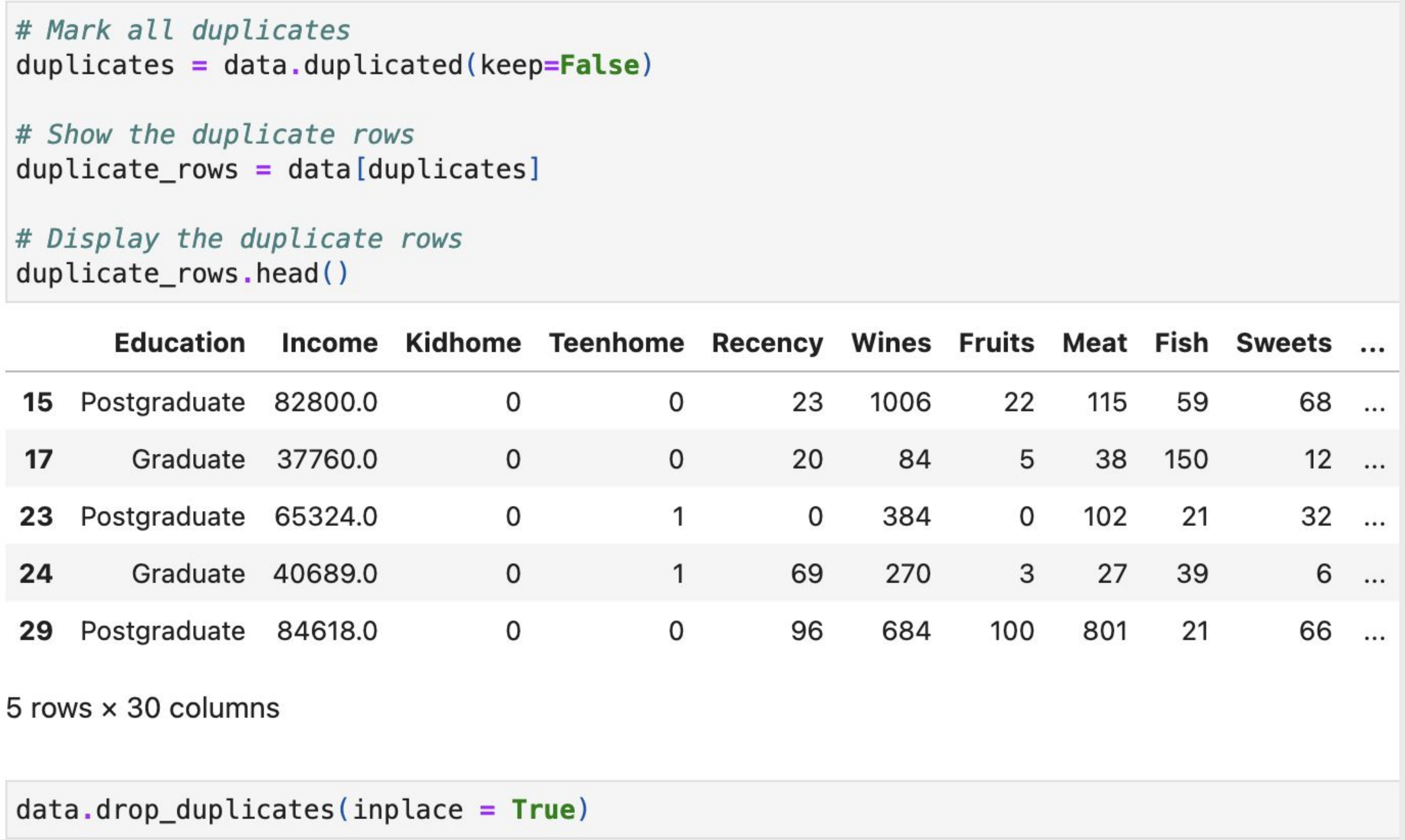
Validation

- Hold-out sampling
  - Split the dataset into training and testing to generalize new data
  - Allows for the evaluation to be unbiased and avoid overfitting



Code Milestones

- Dropped NA values, outliers, unimportant features of modeling, and duplicates
- Handling categorical features
  - Label Encoding
- Scaling data to ensure more accurate results (clusters)
- Unscaled clusters to interpret the data in original units



```
# Mark all duplicates
duplicates = data.duplicated(keep=False)

# Show the duplicate rows
duplicate_rows = data[duplicates]

# Display the duplicate rows
duplicate_rows.head()
```

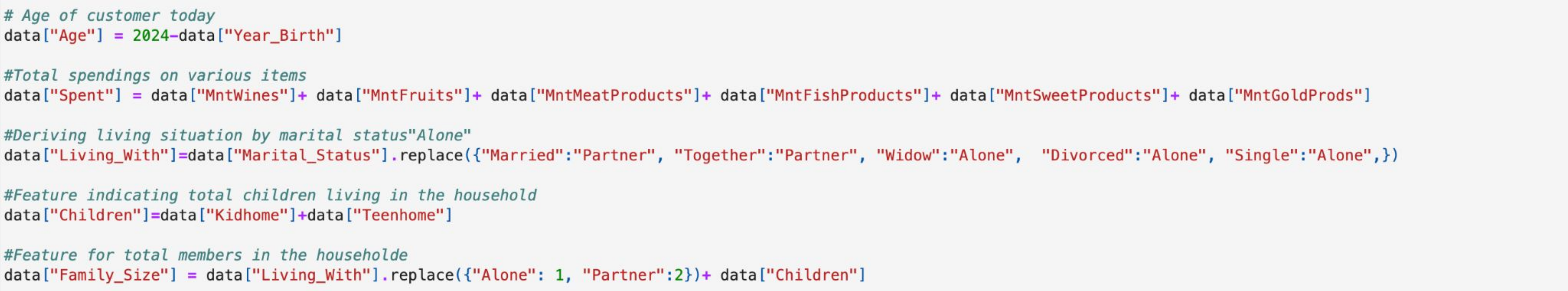
	Education	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	...
15	Postgraduate	82800.0	0	0	23	1006	22	115	59	68	...
17	Graduate	37760.0	0	0	20	84	5	38	150	12	...
23	Postgraduate	65324.0	0	1	0	384	0	102	21	32	...
24	Graduate	40689.0	0	1	69	270	3	27	39	6	...
29	Postgraduate	84618.0	0	0	96	684	100	801	21	66	...

5 rows × 30 columns

```
data.drop_duplicates(inplace = True)
```

Figure 1) Snippet of code dropping duplicates

Feature Engineering



```
# Age of customer today
data["Age"] = 2024-data["Year_Birth"]

#Total spendings on various items
data["Spent"] = data["MntWines"]+ data["MntFruits"]+ data["MntMeatProducts"]+ data["MntFishProducts"]+ data["MntSweetProducts"]+ data["MntGoldProds"]

#Deriving living situation by marital status"Alone"
data["Living_With"]=data["Marital_Status"].replace({"Married":"Partner", "Together":"Partner", "Widow":"Alone",  "Divorced":"Alone", "Single":"Alone",})

#Feature indicating total children living in the household
data["Children"]=data["Kidhome"]+data["Teenhome"]

#Feature for total members in the householde
data["Family_Size"] = data["Living_With"].replace({"Alone": 1, "Partner":2})+ data["Children"]
```

Figure 1) Snippet of Feature Engineering Process

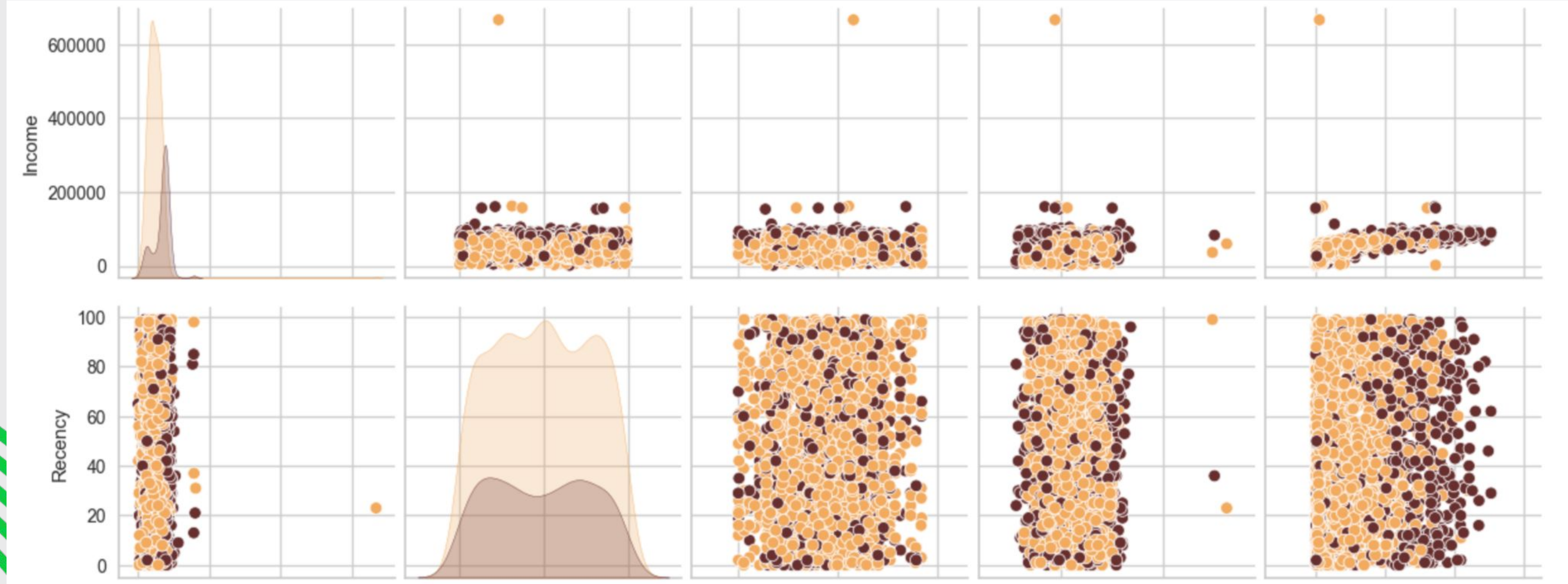


Figure 2) Snippet of plotted features

K-Means

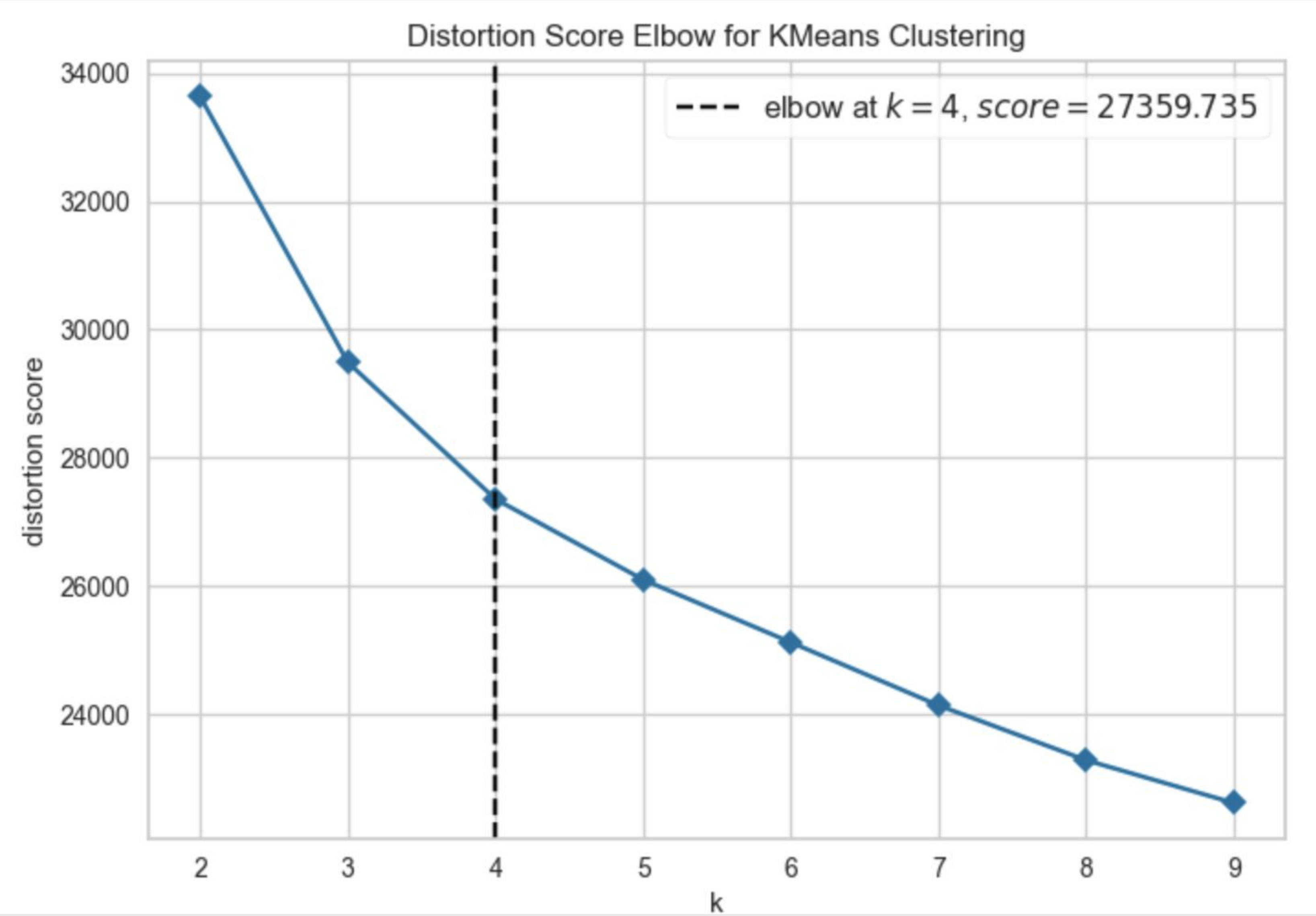


Figure 1) Distortion Score Elbow for K-Means Clustering

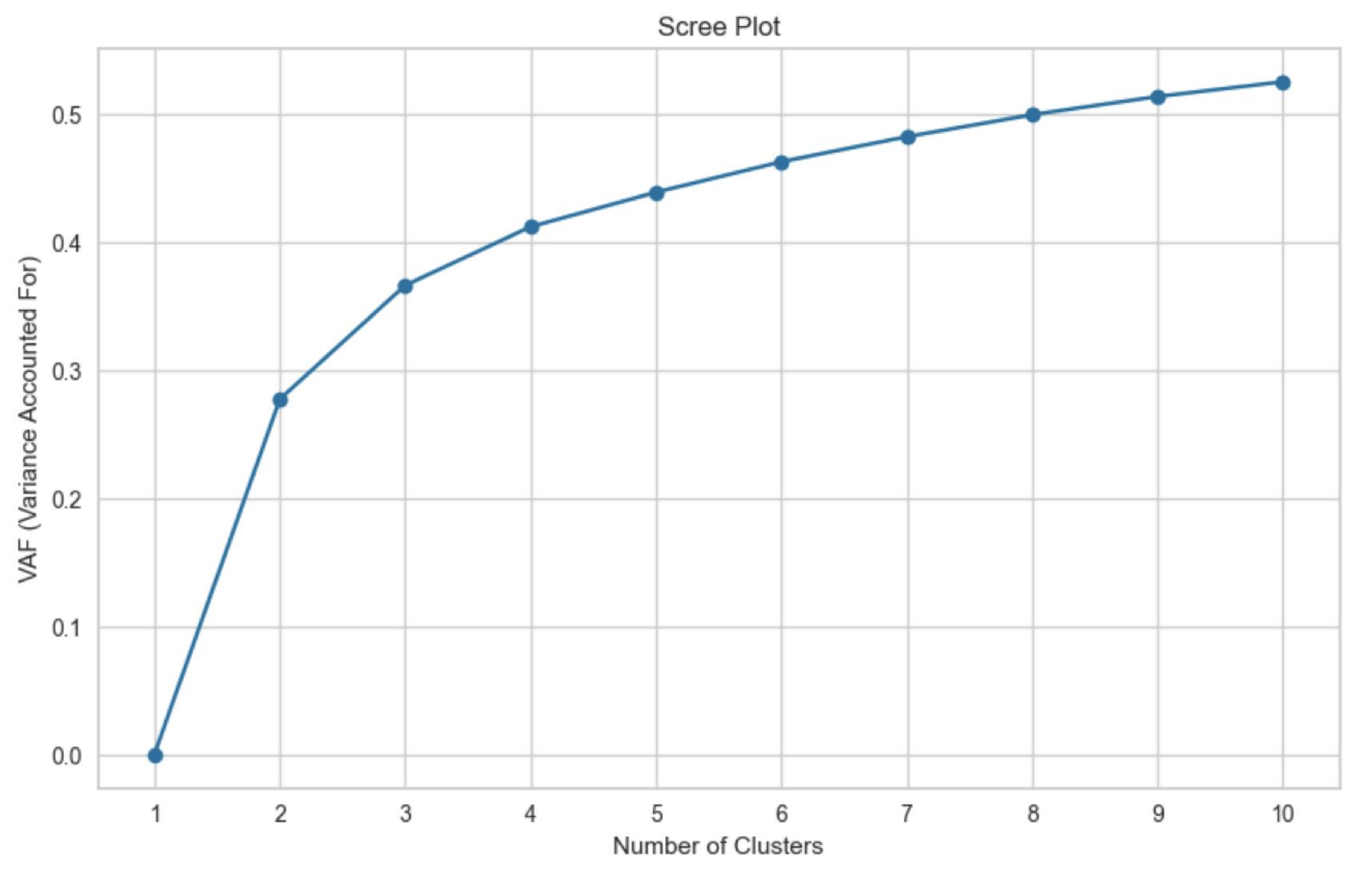
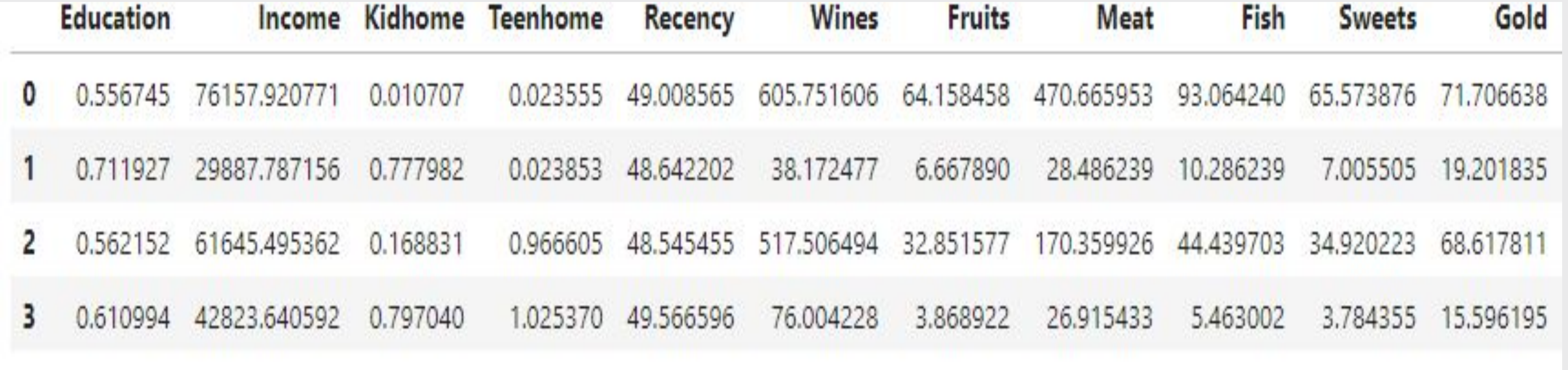


Figure 2) Scree Plot

Number of Clusters: 4

K-Means

- Grouping customers based on their characteristics as shoppers and demographic info
- Helps us to target marketing campaigns and understand what makes certain customers

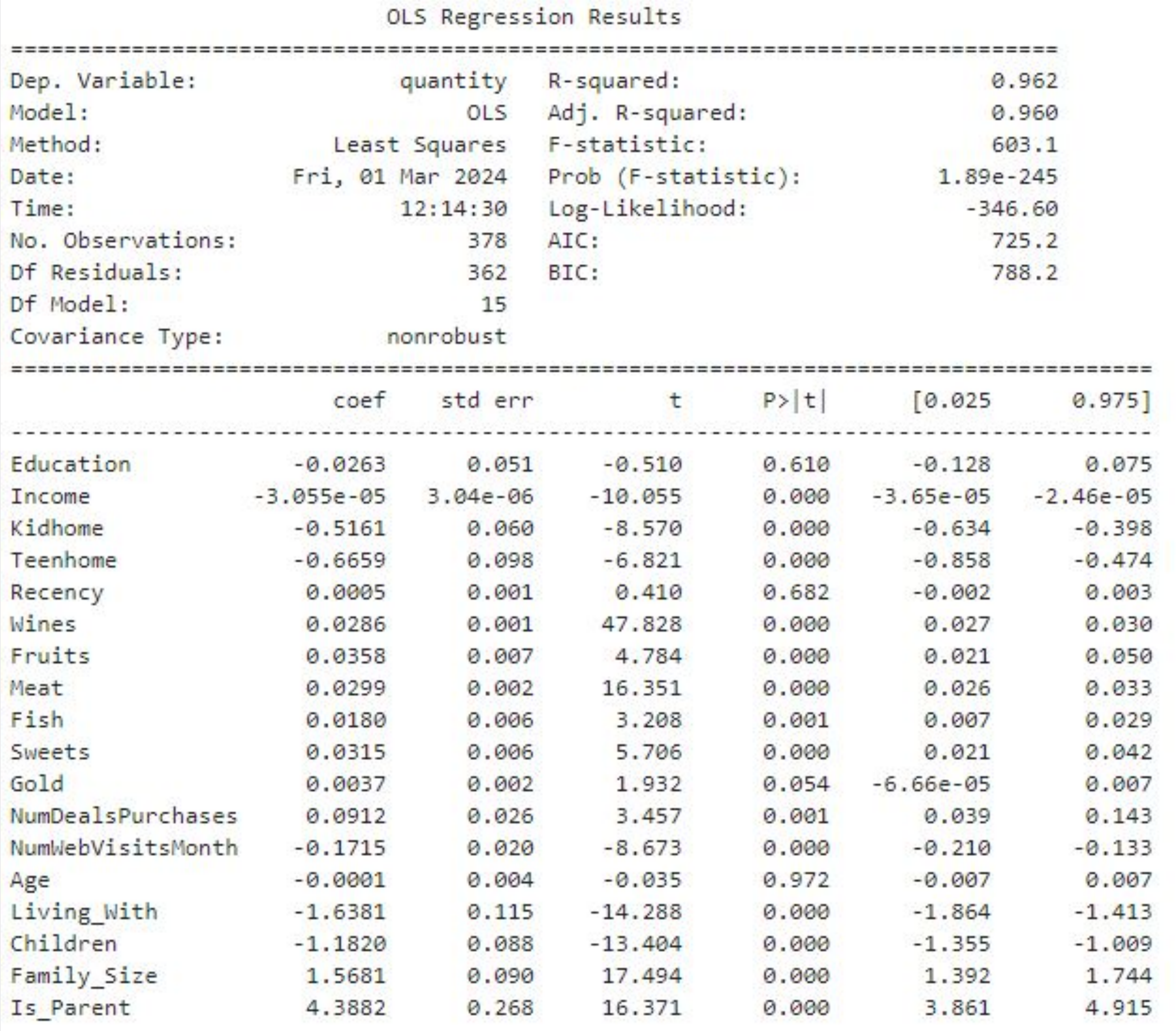


	Education	Income	Kidhome	Teenhome	Recency	Wines	Fruits	Meat	Fish	Sweets	Gold
0	0.556745	76157.920771	0.010707	0.023555	49.008565	605.751606	64.158458	470.665953	93.064240	65.573876	71.706638
1	0.711927	29887.787156	0.777982	0.023853	48.642202	38.172477	6.667890	28.486239	10.286239	7.005505	19.201835
2	0.562152	61645.495362	0.168831	0.966605	48.545455	517.506494	32.851577	170.359926	44.439703	34.920223	68.617811
3	0.610994	42823.640592	0.797040	1.025370	49.566596	76.004228	3.868922	26.915433	5.463002	3.784355	15.596195

Figure 1) K Means Centroids Dataframe

Linear Regression

- Ran one regression for each cluster (4 total) on newly created “quantity” variable
- Main focus on what products for each cluster lead to more purchases
- Analyzing coefficients can help us to understand customer purchase behaviors per cluster



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               quantity   R-squared:                       0.962
Model:                            OLS   Adj. R-squared:                  0.960
Method:                 Least Squares   F-statistic:                     603.1
Date:                Fri, 01 Mar 2024   Prob (F-statistic):          1.89e-245
Time:                        12:14:30   Log-Likelihood:                -346.60
No. Observations:                 378   AIC:                             725.2
Df Residuals:                     362   BIC:                             788.2
Df Model:                          15
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Education               -0.0263      0.051     -0.510      0.610      -0.128       0.075
Income               -3.055e-05   3.04e-06    -10.055      0.000   -3.65e-05   -2.46e-05
Kidhome                 -0.5161      0.060     -8.570      0.000      -0.634      -0.398
Teenhome                -0.6659      0.098     -6.821      0.000      -0.858      -0.474
Recency                  0.0005      0.001      0.410      0.682      -0.002       0.003
Wines                    0.0286      0.001     47.828      0.000       0.027       0.030
Fruits                   0.0358      0.007      4.784      0.000       0.021       0.050
Meat                     0.0299      0.002     16.351      0.000       0.026       0.033
Fish                     0.0180      0.006      3.208      0.001       0.007       0.029
Sweets                   0.0315      0.006      5.706      0.000       0.021       0.042
Gold                     0.0037      0.002      1.932      0.054   -6.66e-05       0.007
NumDealsPurchases        0.0912      0.026      3.457      0.001       0.039       0.143
NumWebVisitsMonth       -0.1715      0.020     -8.673      0.000      -0.210      -0.133
Age                     -0.0001      0.004     -0.035      0.972      -0.007       0.007
Living_With             -1.6381      0.115    -14.288      0.000      -1.864      -1.413
Children                -1.1820      0.088    -13.404      0.000      -1.355      -1.009
Family_Size              1.5681      0.090     17.494      0.000       1.392       1.744
Is_Parent                4.3882      0.268     16.371      0.000       3.861       4.915
```

Figure 1) Output of Regression

Summary Analysis

03

Elaborate on what you want to discuss.

Cluster 1: The Affluent Connoisseurs

Characteristics: Highest income group with significant spending on wines, fruits, and meat, indicating a preference for luxury or gourmet products. Lowest in the household with children and teenagers, suggesting possibly older demographics or empty nesters.

Interpretation of Features: High income and low family obligations allow for significant discretionary spending, particularly on premium goods. Their low recency indicates recent interactions with the brand, suggesting loyalty or ongoing engagement.

Strategic Insights: This segment values quality and is less price-sensitive. Tailored marketing emphasizing quality, exclusivity, and premium offerings could resonate well. Loyalty programs or exclusive events could further enhance their engagement.

Profile Summary: Affluent, likely older customers with a taste for luxury, highly engaged, and with significant purchasing power.

Cluster 1: The Affluent Connoisseurs

Product Preferences: High spending on Wines, Fruits, Meat, and Fish, indicating a strong preference for premium and gourmet items.

Marketing Plan Enhancements:

Curated Wine Selections: Offer exclusive wine subscriptions featuring rare and premium wines, tailored to their sophisticated taste.

Gourmet Hampers: Introduce luxury hampers that include a selection of premium meats, exotic fruits, and high-quality fish – perfect for the gourmet enthusiast.

Sweet and Gold Combos: Create exclusive gift sets combining artisan sweets with gold-themed items, catering to their taste for luxury and quality.

Cluster 2: Budget-Focused Digital-Savvy Young Parents

Characteristics: Lower income and high Kidhome values indicate younger families on a budget. Their spending is much lower across all categories, especially on luxury items like wines and meats. They also have low spending on sweets.

Interpretation of Features: The financial constraints and family focus suggest a prioritization of essential purchases over luxury. The high Kidhome value implies marketing efforts should focus on family-friendly products and value deals.

Strategic Insights: This group may respond well to discounts and value packs. Marketing strategies that highlight affordability, family packages, and essential goods could be effective.

Profile Summary: Younger families managing on a tighter budget, likely to be receptive to offers that maximize value and cater to family needs.

Cluster 2: Budget-Focused Digital-Savvy Young Parents

Product Preferences: Lower spending overall, with modest purchases in basic categories like Fruits and Meats.

Marketing Plan Enhancements:

Value Packs: Promote family-friendly value packs that offer great deals on fruits and meats, ensuring affordability without compromising on health.

Educational Deals: Highlight products that combine value with educational content for children, such as DIY fruit snack kits or interactive meal prep packages.

Deal Alerts: Implement targeted deal alerts for budget-friendly items in their preferred categories, ensuring they never miss out on savings.

Cluster 3: Upper Mid-Level Affluents

Characteristics: Mid-level income with balanced spending across categories. Notably higher engagement in web purchases and a mixed family structure (children and teenagers).

Interpretation of Features: A comfortable income allows for discretionary spending, but choices may be more balanced between necessity and luxury. Their shopping behavior and family structure suggest a versatile consumer group that balances quality with cost.

Strategic Insights: This segment might appreciate a blend of quality and value. Offering exclusive online deals or highlighting quality products with competitive pricing could attract this group. Content that appeals to both parents and older children could increase engagement.

Profile Summary: Financially comfortable, active online shoppers with diverse needs, responding well to quality, value, and convenience.

Cluster 3: Upper Mid-Level Affluents

Product Preferences: Significant spending on Wines and moderate spending across Fruits, Meat, and Fish, indicating a balance of quality and value.

Marketing Plan Enhancements:

Exclusive Online Wine Tastings: Invite to virtual wine tasting sessions featuring selections from their preferred wine categories.

Balanced Meal Kits: Offer meal kits that provide a balanced mix of quality and value across their preferred food categories, emphasizing convenience and taste.

Sustainability Highlight: Focus on products that offer a blend of sustainability and quality, particularly in their preferred categories, appealing to their values and preferences.

Cluster 4: The Economical Engagers

Characteristics: Lower-middle income with the highest number of children and teenagers, indicating possibly larger families. Their spending is focused more on necessities with modest engagement in all purchase types.

Interpretation of Features: Budget-conscious due to larger family size but engages with the brand across different channels. Their spending pattern suggests a focus on practical and essential items over luxury goods.

Strategic Insights: Effective communication could highlight bulk purchase discounts, loyalty programs, and practical products that offer good value for money. Engaging them with educational content on budgeting or value maximization could enhance brand loyalty.

Profile Summary: Larger families with budget constraints, engaged but selective in their purchasing decisions, likely to appreciate value and practicality.

Cluster 4: The Economical Engagers

Product Preferences: Generally lower spending, with a slight preference for more economical options across all categories.

Marketing Plan Enhancements:

Bulk and Economy Buys: Emphasize bulk purchase options and economy packs in essential categories like meats and fruits, maximizing value.

Practical Rewards: Offer rewards or points for purchases in their preferred categories, which can be redeemed for essential household items or discounts.

Value-Focused Workshops: Host workshops focused on economical meal planning and cooking, using products from their preferred categories to maximize engagement.

Implications

- Integrate specific product recommendations into marketing plans for each cluster → more tailored marketing strategy to the distinct preferences and needs of each customer segment
- Enhance customer satisfaction, loyalty, and sales performance across all segments
- Drive future product development for specific customer segments

THANK YOU

Write a closing statement or call-to-action here.